CSEP 517
Natural Language Processing
Autumn 2015

Introduction
Yejin Choi

Slides adapted from Dan Klein, Luke Zettlemoyer

People & Page

- Page:
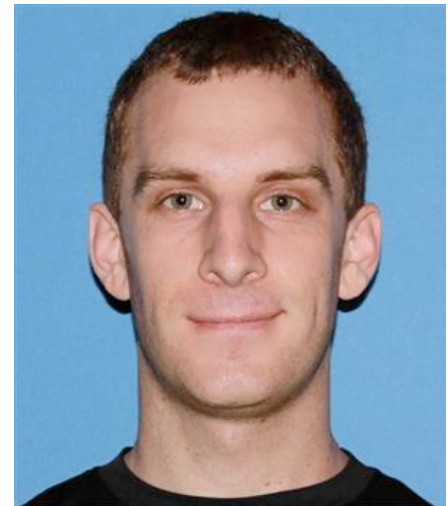 - http://courses.cs.washington.edu/courses/csep517/15au/
- People:
  - Instructor: Yejin Choi
  - TAs: Ignacio "nacho" cano & James Ferguson

Ignacio Agustin
Cano

James F Ferguson

Study Aid

What is NLP?

- Fundamental goal: *deep* understand of *broad* language
  - Not just string processing or keyword matching

- End systems that we want to build:
  - Simple: spelling correction, text categorization…
  - Complex: speech recognition, machine translation, information extraction, sentiment analysis, question answering…
  - Unknown: human-level comprehension (is this just NLP?)

Machine Translation

"Il est impossible aux journalistes de rentrer dans les régions tibétaines"

Bruno Philip, correspondant du "Monde" en Chine, estime que les journalistes de l'AFP qui ont été expulsés de la province tibétaine du Qinghai "n'étaient pas dans l'illégalité".

Les faits Le dalaï-lama dénonce l'"enfer" imposé au Tibet depuis sa fuite, en 1959
Vidéo Anniversaire de la rébellion tibétaine : la Chine sur ses gardes

"It is impossible for journalists to enter Tibetan areas"

Philip Bruno, correspondent for "World" in China, said that journalists of the AFP who have been deported from the Tibetan province of Qinghai "were not illegal."

Facts The Dalai Lama denounces the "hell" imposed since he fled Tibet in 1959
Video Anniversary of the Tibetan rebellion: China on guard

- Translate text from one language to another
- Recombines fragments of example translations
- Challenges:
  - What fragments? [learning to translate]
  - How to make efficient? [fast translation search]
  - Fluency (second half of this class) vs fidelity (later)

2013 Google Translate: French

2013 Google Translate: Russian

Поиск

ENG RUS PT ITA

Например: Большой Кавказ

Мир | Наука | Общество | Здоровье | Красота |

■ Новости

20:09
В Шри-Ланке хотели перевезти золото в желудках

20:00
Выходец из России может получить "Нобеля" по химии

19:46
В США установили стандарты торговли оружием

19:35
Директор Эрмитажа: Обыски нанесли ущерб музею

19:25
Мозгу ребенка полезен послеобеденный сон

19:24
Роликом с водителями-детьми заинтересовалась петербургская полиция

19:15
К Марсу приближается "комета века"

18:55
Выявлено более 160 нарушений на судостроительных предприятиях

18:44
Астахов назначен на новый срок в Европейской сети детских омбудсменов

■ Главное

"Обиженные люди работают, а иностр к нам не поедут"

25.09.2013 19:48

Ректор "Бауманки" Анатолий "Правде.Ру", какие шаги надо чиновникам и ученым в связи реформе РАН.

Фотосессия

Наводнение в Индии: 40 жителей эвакуированы

Найроби. Газета The Independ "Уэстгейт" во время захвата.

■ Мир

Поиск

ENG RUS PT ITA

For example, the Greater Caucasus

World | Science | Society | Health | Beauty | Regions | Photo | Video | Forums | archive

■ News

20:09
In Sri Lanka, wanted to carry the gold in the stomachs

20:00
A native of Russia can get the "Nobel" in Chemistry

19:46
In the United States set the standard arms trade

19:35
Director of the Hermitage: The searches have damaged the museum

19:25
The child's brain is useful afternoon nap

19:24
The roller with the drivers, children become interested in the St. Petersburg Police

19:15
To Mars is approaching "comet of the century"

18:55
There are over 160 violations at shipyards

18:44
Astakhov appointed for a new term in the European Network of Ombudsmen for children

■ Point

"Mentally ill people are working, and foreign scholars to us will not go"

25/09/2013 19:48

The Rector, "Bauman" Anatoly Alexandrov told with "Pravda.Ru" what steps need to be taken to officials and scientists in connection with the adoption of the law on the reform of the RAS.

Photoshoot

2466 photos

World through the lens: September 25,

In Kenya - mourning for the victims of the terrorist attack in Nairobi. The newspaper The Independent said about the people who were at the mall, "Westgate" during capture.

Expert: The poorer the society is, the more scandals due to copyright

09/25/2013 20:04

Why Russians are greedy for free, and do not like to pay for downloading movies and music, with "Pravda.Ru" said the head of Liveinternet German Klimenko.

Putin met environmentalists "Greenpeace" trying to grab the platform

25/09/2013 14:39

President of Russia, speaking at the International Arctic Forum in Salekhard, spoke about the ecology of Greenpeace, staged on a platform of "Prirazlomnaja."

Expert: It is necessary to encourage participation in the election, rather than returning the column "against all"

09/25/2013 13:27

Political scientist and philosopher, Professor Oleg Matveychev HSE commented with "Pravda.Ru" Valentina Matviyenko offer to return to the ballot line "against all."

The British newspaper described the heroes and victims in Nairobi

25/09/2013 10:27

■ World   ■ Policy   ■ Economy

Jeopardy! World Champion

US Cities: Its largest airport is named for a World War II hero; its second largest, for a World War II battle.

Information Extraction

- Unstructured text to database entries

New York Times Co. named Russell T. Lewis, 45, president and general manager of its flagship New York Times newspaper, responsible for all business-side activities. He was executive vice president and deputy general manager. He succeeds Lance R. Primis, who in September was named president and chief operating officer of the parent.

Person	Company	Post	State
Russell T. Lewis	New York Times newspaper	president and general manager	start
Russell T. Lewis	New York Times newspaper	executive vice president	end
Lance R. Primis	New York Times Co.	president and CEO	start

- SOTA: perhaps 80% accuracy for multi-sentence temples, 90%+ for single easy fields
- But remember: information is redundant!

Knowledge Graph: "*things not strings*"

Question Answering

- **Question Answering:**
  - More than search
  - Can be really easy: "What's the capital of Wyoming?"
  - Can be harder: "How many US states' capitals are also their largest cities?"
  - Can be open ended: "What are the main issues in the global warming debate?"

- **Natural Language Interaction:**
  - Understand requests and act on them
  - "Make me a reservation for two at Quinn's tonight''

Google — Web Images Groups News Froogle Local **more »**

any US states' capitals are also their largest cities?   Search

Web

Your search - **How many US states' capitals are also their largest cities?** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Google Home - - Business Solutions - About Google

capital of Wyoming: Information From Answers.com
Note: click on a word meaning below to see its connections and related words.
The noun **capital** of **Wyoming** has one meaning: Meaning #1 : the **capital**.
www.answers.com/topic/**capital**-of-**wyoming** - 21k - Cached - Similar pages

Cheyenne: Weather and Much More From Answers.com
Chey·enne (shī-ăn ' , -ĕn ') The **capital** of **Wyoming**, in the southeast part of the state near the Nebraska and Colorado borders.
www.answers.com/topic/cheyenne-**wyoming** - 74k - Cached - Similar pages

Mobile devices can now answer (some or our) questions and execute commands...

Will this Be Part of All Our Home Devices?

https://www.youtube.com/watch?v=KkOCeAtKHIc
https://www.youtube.com/watch?v=qGU-SqUTees

- Why are these appearing now?
- What are fundamental limitations in current art?
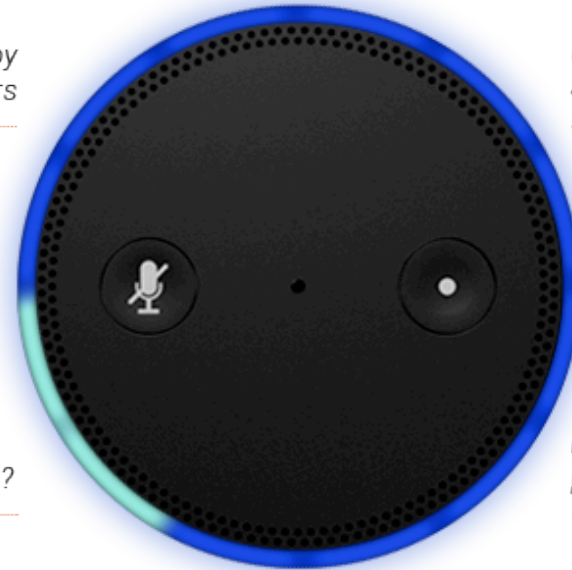
Language Comprehension?

"The rock was still wet. The animal was glistening, like it was still swimming," recalls Hou Xianguang. Hou discovered the unusual fossil while surveying rocks as a paleontology graduate student in 1984, near the Chinese town of Chengjiang. "My teachers always talked about the Burgess Shale animals. It looked like one of them. My hands began to shake." Hou had indeed found a Naraoia like those from Canada. However, Hou's animal was 15 million years older than its Canadian relatives.

It can be inferred that Hou Xianguang's "hands began to shake", because he was:

- (A)   afraid that he might lose the fossil
- (B)   worried about the implications of his finding
- (C)   concerned that he might not get credit for his work
- (D)   uncertain about the authenticity of the fossil
- (E)   excited about the magnitude of his discovery

Speech Recognition

- **Automatic Speech Recognition (ASR)**
  - Audio in, text out
  - SOTA: 0.3% error for digit strings, 5% dictation, 50%+ TV
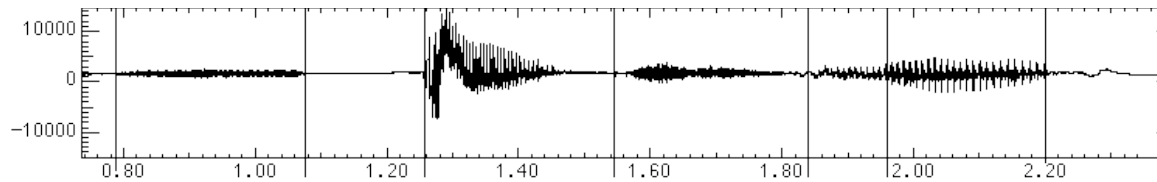
"Speech Lab"

- **Text to Speech (TTS)**
  - Text in, audio out
  - SOTA: totally intelligible (if sometimes unnatural)

Analyzing public opinion, making political forecasts

- Today: In 2012 election, automatic sentiment analysis actually being used to complement traditional methods (surveys, focus groups)
- Past: "Sentiment Analysis" research started in 2002 or so
- Future: Growing research toward computational social science, digital humanities (psychology, communication, literature and more)
- Challenge: Need statistical models for deeper semantic understanding --- subtext, intent, nuanced messages
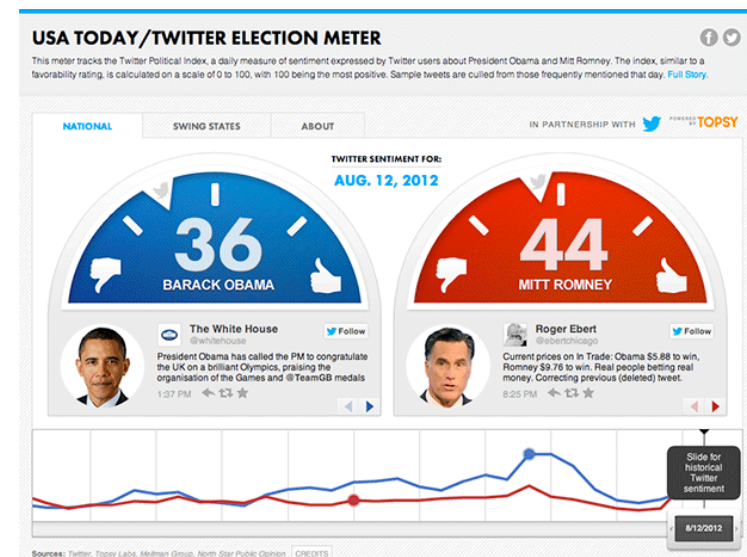
Summarization

- Condensing documents
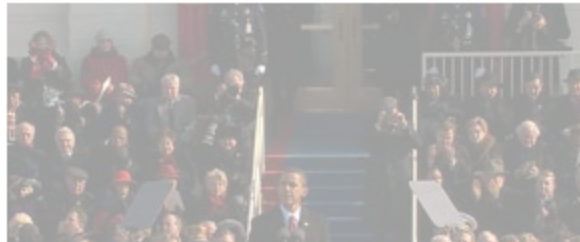 - Single or multiple docs
  - Extractive or synthetic
  - Aggregative or representative

- Very context-dependent!

- An example of analysis with generation

WASHINGTON (CNN) -- President Obama's inaugural address was cooler, more measured and reassuring than that of other presidents making it, perhaps, the right speech for the times.

Some inaugural addresses are known for their soaring, inspirational language. Like John F. Kennedy's in 1961: "Ask not what your country can do for you. Ask what you can do for your country."

Obama's address was less stirring, perhaps, but it was also more candid and down-to-earth.

"Starting today," the new president said, "we must ... begin

STORY HIGHLIGHTS
- Obama's address less stirring than others but more candid, analyst says
- Schneider: At a time of crisis, president must be reassuring
- Country has chosen "hope over fear, unity of purpose over ... discord," Obama said
- Obama's speech was a cool speech, not a hot one, Schneider says

said in his first inaugural in 1933, "The only thing we have to fear is fear itself." Or Bill Clinton, who took office during the economic crisis of the early 1990s. "There is nothing wrong with America that cannot be fixed by what is right with America," Clinton declared at his first inaugural.

President Obama renewed his call for a massive plan to stimulate economic growth.

more photos »

Obama, too, offered reassurance.

"We gather because we have chosen hope over fear, unity of purpose over conflict and discord," Obama said.

Obama's call to unity after decades of political division echoed Abraham Lincoln's first inaugural address in 1861. Even though he delivered it at the onset of a terrible civil war, Lincoln's speech was not a call to battle. It was a call to look beyond the war, toward reconciliation based on what he called "the better angels of our nature."

Some presidents used their inaugural address to set out a bold agenda.

Start-up Summly → Yahoo!

CEO Marissa Mayer announced an update to the app in a blog post, saying, "The new Yahoo! mobile app is also smarter, using Summly's natural-language algorithms and machine learning to deliver quick story summaries. We acquired Summly less than a month ago, and we're thrilled to introduce this game-changing technology in our first mobile application."

Launched 2011, Acquired 2013 for $30M

Can a robot write news?

Despite an expected dip in profit, analysts are generally optimistic about **Steelcase** as it prepares to reports its third-quarter earnings on Monday, December 22, 2014. The consensus earnings per share estimate is 26 cents per share.

The consensus estimate remains unchanged over the past month, but it has decreased from three months ago when it was 27 cents. Analysts are expecting earnings of 85 cents per share for the fiscal year. Revenue is projected to be 5% above the year-earlier total of $784.8 million at $826.1 million for the quarter. For the year, revenue is projected to come in at $3.11 billion.

The company has seen revenue grow for three quarters straight. The less than a percent revenue increase brought the figure up to $786.7 million in the most recent quarter. Looking back further, revenue increased 8% in the first quarter from the year earlier and 8% in the fourth quarter.

The majority of analysts (100%) rate Steelcase as a buy. This compares favorably to the analyst ratings of three similar companies, which average 57% buys. Both analysts rate Steelcase as a buy.

Steelcase is a designer, marketer and manufacturer of office furniture. Other companies in the furniture and fixtures industry with upcoming earnings release dates include: HNI and Knoll.

Writer-bots for earthquake & financial reports

Some of the formulaic news articles are now written by computers.

•Definitely far from "Op-ed"
•Can we make the generation engine statistically learned rather than engineered?

Language and Vision

"Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description. This is a problem of overwhelming difficulty, relying as it does on finding solutions to both vision and language and then integrating them. I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers"

-- David Stork (HAL's Legacy, 2001) on A. Rosenfeld's vision
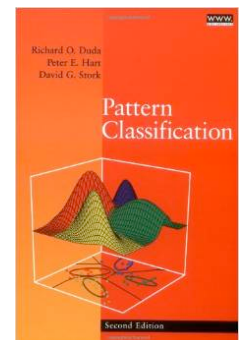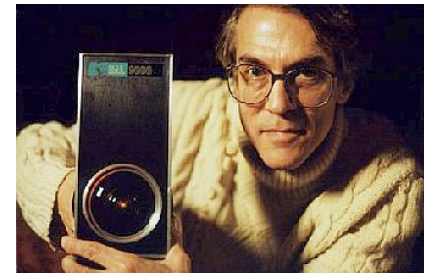
What begins to work (e.g., Kuznetsova et al. 2014)

The flower was so vivid and attractive.

Blue flowers are running rampant in my garden.

We sometimes do well: 1 out of 4 times, machine captions were preferred over the original Flickr captions:

Spring in a white dress.

Blue flowers have no scent. Small white flowers have no idea what they are.

Scenes around the lake on my bike ride.

This horse walking along the road as we drove by.

But many challenges remain
(better examples of when things go awry)

The couch is definitely bigger than it looks in this photo.

Yellow ball suspended in water.

Incorrect Object Recognition

Incorrect Scene Matching

Incorrect Composition

My cat laying in my duffel bag.

A high chair in the trees.

NLP History: pre-statistics

- **(1) Colorless green ideas sleep furiously.**
- **(2) Furiously sleep ideas green colorless**
  - It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) had ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally "remote" from English. Yet (1), though nonsensical, is grammatical, while (2) is not." (Chomsky 1957)
- **70s and 80s: more linguistic focus**
  - Emphasis on deeper models, syntax and semantics
  - Toy domains / manually engineered systems
  - Weak empirical evaluation

NLP: machine learning and empiricism

> "Whenever I fire a linguist our system performance improves." –Jelinek, 1988

- 1990s: Empirical Revolution
  - Corpus-based methods produce the first widely used tools
  - Deep linguistic analysis often traded for robust approximations
  - *Empirical evaluation* is essential
- 2000s: Richer linguistic representations used in statistical approaches, scale to more data!
- 2010s: you decide!

What is Nearby NLP?

- ## Computational Linguistics
  - Using computational methods to learn more about how language works
  - We end up doing this and using it
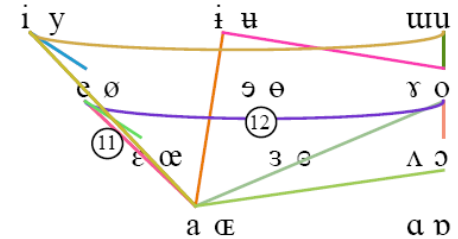
- ## Cognitive Science
  - Figuring out how the human brain works
  - Includes the bits that do language
  - Humans: the only working NLP prototype!
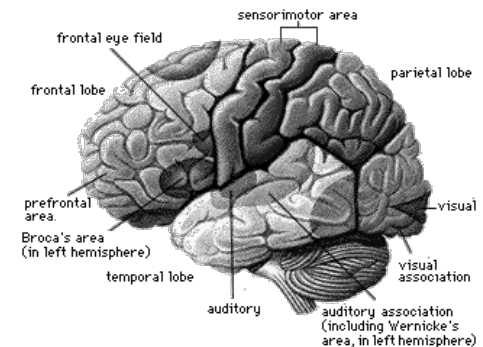
- ## Speech?
  - Mapping audio signals to text
  - Traditionally separate from NLP, converging?
  - Two components: acoustic models and language models
  - Language models in the domain of stat NLP

Problem: Ambiguities

- Headlines:
  - Enraged Cow Injures Farmer with Ax
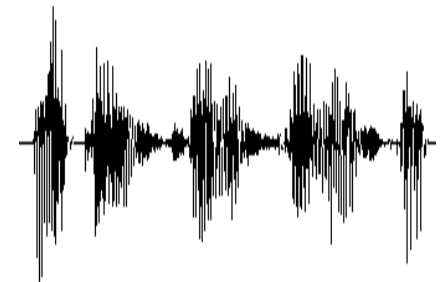 - Ban on Nude Dancing on Governor's Desk
  - Teacher Strikes Idle Kids
  - Hospitals Are Sued by 7 Foot Doctors
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
- Why are these funny?

Syntactic Analysis
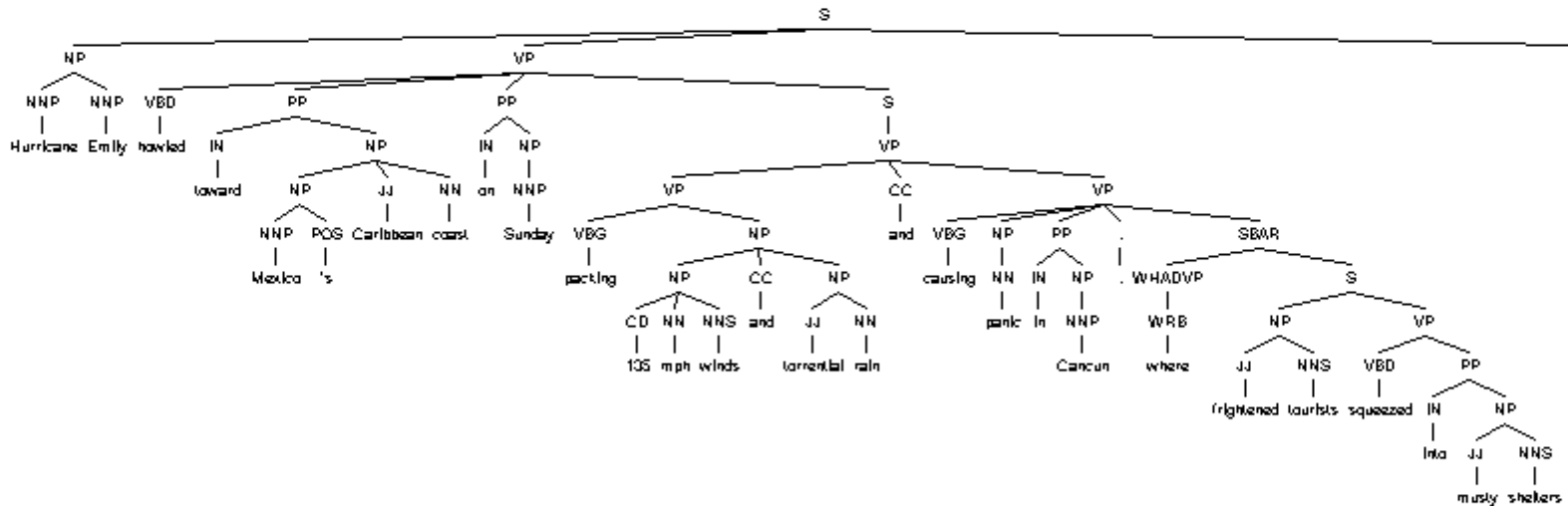
Hurricane Emily howled toward Mexico 's Caribbean coast on Sunday packing 135 mph winds and torrential rain and causing panic in Cancun , where frightened tourists squeezed into musty shelters .

- SOTA: ~90% accurate for many languages when given many training examples, some progress in analyzing languages given few or no examples

Semantic Ambiguity

At last, a computer that understands you like your mother.

- **Direct Meanings:**
  - It understands you like your mother (does) [presumably well]
  - It understands (that) you like your mother
  - It understands you like (it understands) your mother
- **But there are other possibilities, e.g. mother could mean:**
  - a woman who has given birth to a child
  - a stringy slimy substance consisting of yeast cells and bacteria; is added to cider or wine to produce vinegar
- **Context matters, e.g. what if previous sentence was:**
  - Wow, Amazon predicted that you would need to order a big batch of new vinegar brewing ingredients. ☺
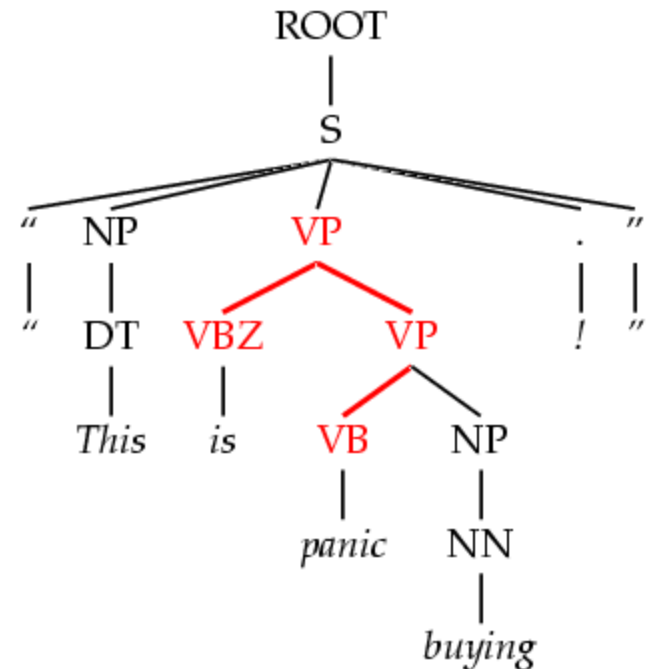
[Example from L. Lee]

Dark Ambiguities

- *Dark ambiguities*: most structurally permitted analyses are so bad that you can't get your mind to produce them

This analysis corresponds to the correct parse of

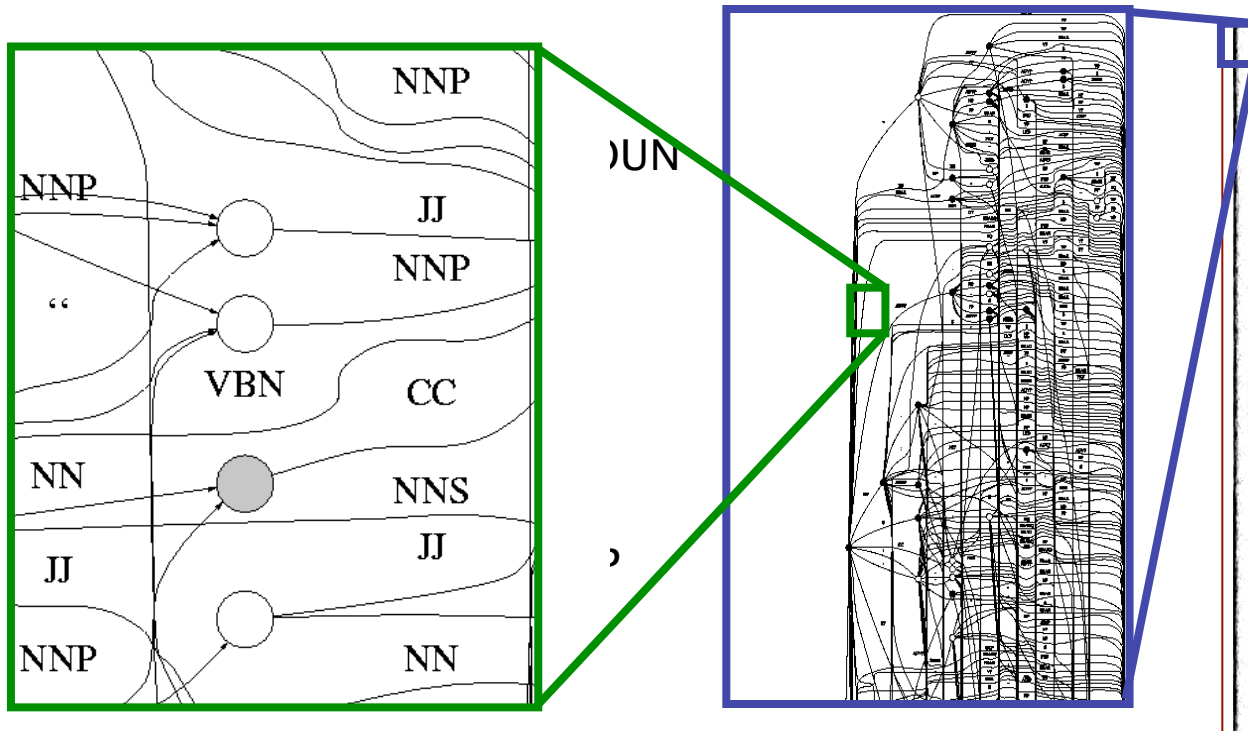"*This will panic buyers ! *"

- Unknown words and new usages
- Solution: We need mechanisms to focus attention on the best ones, probabilistic techniques do this

Problem: Scale

- People *did* know that language was ambiguous!
  - …but they hoped that all interpretations would be "good" ones (or ruled out pragmatically)
  - …they didn't realize how bad it would be

Corpora
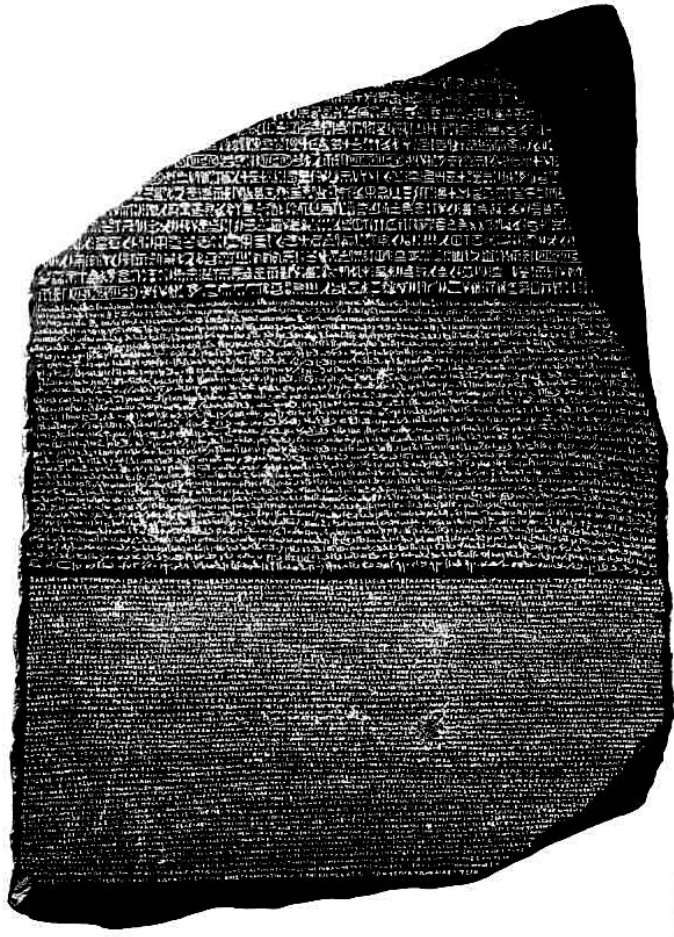
- **A corpus is a collection of text**
  - Often annotated in some way
  - Sometimes just lots of text
  - Balanced vs. uniform corpora

- **Examples**
  - Newswire collections: 500M+ words
  - Brown corpus: 1M words of tagged "balanced" text
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of aligned French / English sentences
  - The Web: billions of words of who knows what

Problem: Sparsity

- ## However: sparsity is always a problem
  - ### New unigram (word), bigram (word pair)
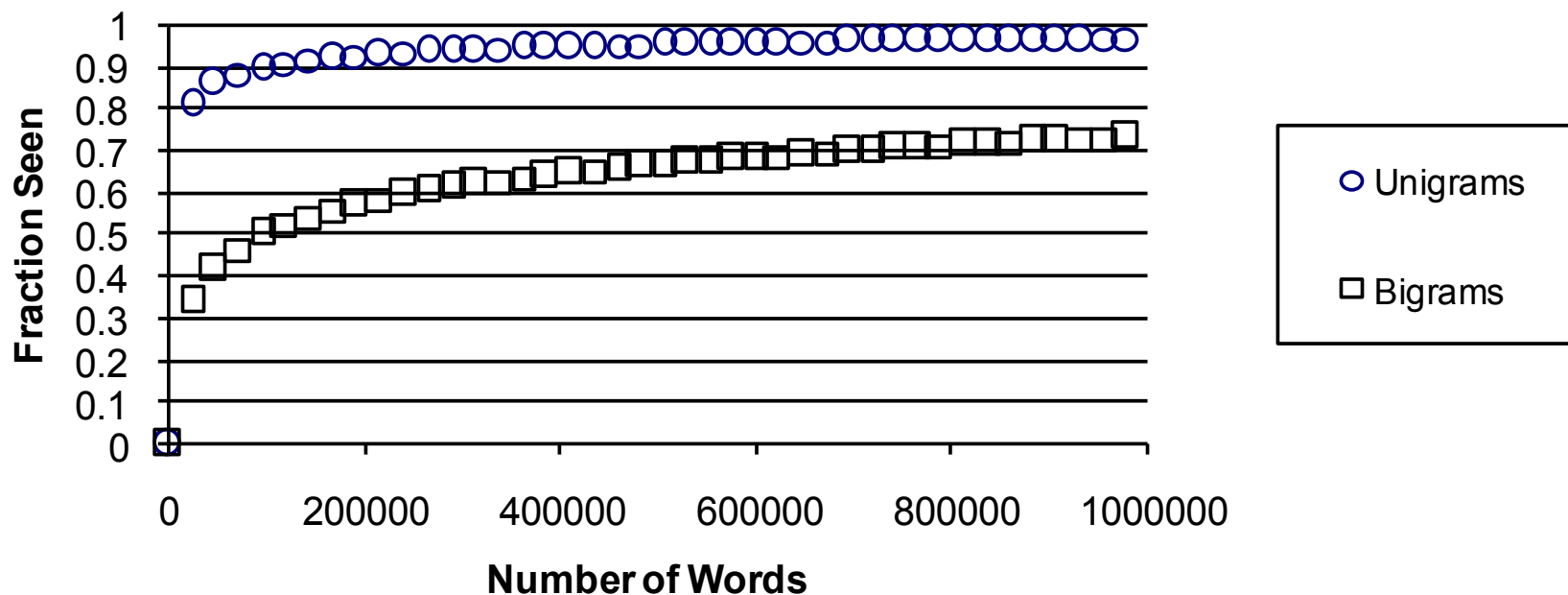
Outline of Topics

| Week	Dates	Topics & Lecture Slides	Notes (Required)	Textbook	Supplementary Readings
1	Oct 6	Introduction ; Language Models (LM) [Slides]	LM Notes	J&M 4.1-4; M&S 6	[Large LMs] [Berkeley LM]
2	Oct 13	**Sequences:** Language Models and Smoothing; Hidden Markov Models (HMMs) [Slides]	HMM Notes	J&M 4.5-7; M&S 6	[Smoothing]
3	Oct 20	Hidden Markov Models (HMMs) [Slides]& Part-Of-Speech Tagging [Slides]		J&M 5.1-5.3; 6.1-6.4; M&S 9, 10.1-10.3	[TnT Tagger] [Stanford Tagger] [SOTA POS]
4	Oct 27	**Trees:** Probabilistic Context Free Grammars (PCFG) and Parsing [Slides]	PCFG Notes, Lexicalized PCFGs	J&M 13-14; M&S 11-12	[Syntax Intro] [Incremental] [Best First] [A* Parsing] [Lexicalized] [Unlexicalized] [Split Merge]
5	Nov 3	More Parsing [Slides]; Expectation Maximization (EM)[Slides]	EM Notes, Forward-backward, Inside-outside	J&M 6.5; M&S 9.3-4; 11.3-4	[Semi-supervised Naive Bayes] [EM Tutorial] [EM for Feature-Rich]
6	Nov 10	**Machine Translation (MT):** Word Alignment [Slides]	IBM Models 1 and 2	J&M 25.1-6; M&S 13	[IBM Models] [HMM Model] [MERT Training]
7	Nov 17	Phrase-based MT [Slides]; Syntax-based MT [Slides I] [II]	Phrase-based Notes	J&M 25.6-10; M&S 13	[SCFG Tutorial] [Hiero] [Tree-to-String] [Tree-to-Tree]
8	Nov 24	**Log-Linear / Feature-Rich Models:** Conditional Random Fields (CRFs) [Slides]	Log-linear models CRF Notes	J&M 6.6-6.8; M&S 16.2-16.3	[MaxExt] [CRF Tutorial] [CRF LM] [CRF Parsing]
9	Dec 1	**Knowledge & Semantic Relations:** Information Extraction; Entailment; [Slides]		J&M 22	[Entailment Graphs] [Paraphrasing w/ MT] [Paraphrasing and Entailment]
10	Dec 8	**Neural Models**			

Course Details

- Books (recommended but required):
  - Jurafsky and Martin, Speech and Language Processing, 2nd Edition (not 1st)
  - Manning and Schuetze, Foundations of Statistical NLP
- Prerequisites:
  - CSE 421 (Algorithms) or equivalent
  - Some exposure to dynamic programming and probability helpful
  - Strong programming
  - There will be a lot of math and programming
- Work and Grading --- TBD!:
  - 3~4 programming-based homeworks (60%), a final mini-project (20%), non-programming assignments (15%) and course/discussion board participation (5%). No midterm or final exam.
- Contact: see website for details
  - Class participation is expected and appreciated!!!
  - Email is great, but please use the message board when possible (we monitor it closely)
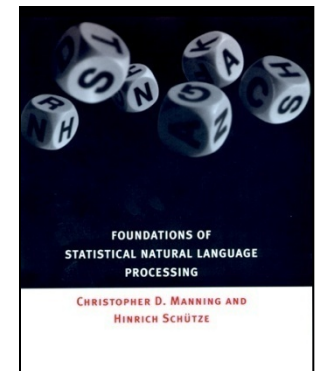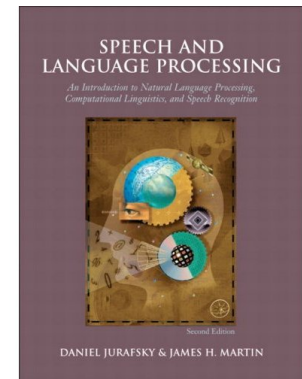
Assignments

Programming-based Assignments (code + write-up)

- Assignment 1: Language Models (Due Mon Oct 19, 5pm)
- Assignment 2: HMMs
- Assignment 3: TBD
- Assignment 4: TBD

(You may choose 3~4 assignments above depending on your plan for the final mini-project.)

Non-programming-based Assignments (10% grade)

- Lightweight 2 written homework assignments

Course participation (10% grade)

Final Mini-project

- Yay / Nay?

Useful Pointers

- **ACL Anthology (http://www.aclweb.org/anthology/)**
  - All NLP conference and journal papers, free of charge

ACL Anthology
A Digital Archive of Research Papers in Computational Linguistics

Search the Anthology | via Google | via Searchbench @ DFKI | via AAN @ UMich | via Saffr

The ACL Anthology currently hosts over 33,000 papers on the study of computational linguistics and natural language processing. Subscribe to the mailing list to

NEW The beta version of the new ACL Anthology goes live. It will replace this current version of the Anthology as the default version starting 2015 (don't worry we w

NEW Nov 2014: The December issue of *Computational Linguistics* is now available on the Anthology. Also the Proceedings of the 26th Conference on Computation and the Proceedings of the Australasian Language Technology Association Workshop 2014 (ALTA 2014) are now available on the Anthology. The Proceedings of Language Learning (NLP4CALL) have also been made available on the Anthology.

ACL events

CL: Intro FS MT&CL 74-79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 **UPDATED** 14

TACL: 14 13

ACL: Intro 79 80 81 82 83 84* 85 86 87 88 89 90 91 92 93 94 95 96 97* 98* 99 00 01 02 03 04 05 06* 07 08* 09* 10 11 12 13 14

EACL: Intro 83 85 87 89 91 93 95 97* 99 03 06 09 12 14

NAACL: Intro 00* 01 03 04 06* 07 09* 10* 12* 13*

EMNLP: 96 97 98 99 00 01 02 03 04 05 06 07* 08 09 10 11 12* 13 14

CoNLL: 97 98 99 00 01 02 03 04 05 06 07 08 09 10 11 12 13 14

*Sem/ SemEval: 98 01 04 07 10 12 13 14

What is this Class?

- Three aspects to the course:
  - Linguistic Issues
    - What are the range of language phenomena?
    - What are the knowledge sources that let us disambiguate?
    - What representations are appropriate?
    - How do you know what to model and what not to model?
  - Statistical Modeling Methods
    - Increasingly complex model structures
    - Learning and parameter estimation
    - Efficient inference: dynamic programming, search, sampling
  - Engineering Methods
    - Issues of scale
    - Where the theory breaks down (and what to do about it)
- We'll focus on what makes the problems hard, and what works in practice…

Class Requirements and Goals

- **Class requirements**
  - Uses a variety of skills / knowledge:
    - Probability and statistics
    - Basic linguistics background
    - Decent coding skills
  - Most people are probably missing one of the above
  - You will often have to work to fill the gaps

- **Class goals**
  - Learn the issues and techniques of modern NLP
  - Build realistic NLP tools
  - Be able to read current research papers in the field
  - See where the holes in the field still are!

Questions for You

Office hour?

Questionnaire:

[1] academic background

- degree, major, institution, year

- AI, ML courses taken

[2] what do you do?

[3] why learning NLP?

[4] what do you envision in future AI?

- one recent application that surprised you

- one dream application you like to see